LEVEL OF SERVICE FOR AVIATION INFRASTRUCTURE MARKETS?

Dipl.-Ing. E. Grunewald
German Aerospace Center
Air Transport and Airport Research
Lilienthalplatz 7, 38108 Braunschweig, Germany

OVERVIEW

In fact, with respect to air traffic, the present state of on-time performance cannot be considered as satisfactory from a customer's point of view. However, the relatively steady situation seen in the recent years indicates a grown optimum with regard to the actual circumstances. Economically effective incentives for on-time performance improvements are missing in the present resource allocation system. For example, the current principle does neither foresee any internalising of external costs arising from delay nor meet the micro-economic requirements of the airlines by sophisticated infrastructure supply. The new resource allocation procedure described here is based on simple principles, which emphasize the monetary value of punctuality. The infrastructure user should be free to individually choose the appropriate level of service and thus the risk of being impacted by delay, due to occupied infrastructure. Furthermore, the user's individual contribution to network performance is to be balanced. A monetary Bonus-Malus-System is to be abolished. Instead, performance-dependent priorities for the usage of short resources should improve the efficiency of resource usage.

The interim results of an acquisition of expert knowledge support the approach. The majority of experts questioned are in favour of differentiable infrastructure services and emphasize the basic principle of resource allocation, EQUITY, instead of EQUALITY. Simultaneously, the interim results reveal the severe concerns with regard to the obstacles to overcome when changing the system. Now, the challenge is to point up the win-win-situation for users and suppliers, which arises from the implementation of a Service-Level Concept.

1. PUNCTUALITY IN AVIATION

So far, in the context of air traffic the definition of on-time performance is not clear and binding yet. Even the wording needs to be fixed before any discussion takes place. In this context, the wording *on-time* or *punctual* describe the status of meeting a time target. A delay is measured as the difference between actual and scheduled event. Having a delay is only then leading to unpunctuality, if the agreed punctuality window is exceeded. As shown in FIG 1, the same might apply for early arrivals too.

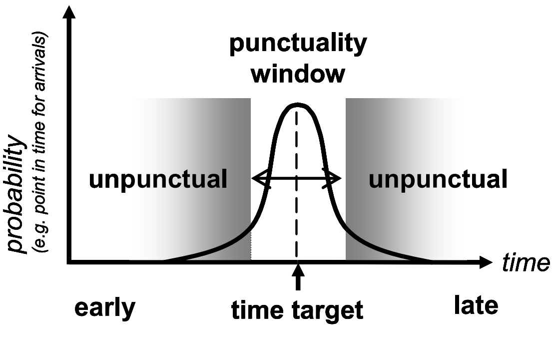

FIG 1. Punctuality declaration

The German Air Traffic Control (DFS) declares those flights as punctual, which do not exceed more than 15 minutes of delay within the responsibility of the air traffic control; whereas, delays before and after the actual flight are not taken into account. The DFS indicator for on-time performance thus refers to a relative measured variable. Absolute measured variables rather record the entire travel chain in the air traffic system: An airplane is punctual, if it does not have any more than 15 minutes of delay in relation to the planned and published (arrival block) time. Also the Advisory Council for Aeronautics Research in Europe (ACARE) is geared to the 15-minutes-window, as it defines its punctuality vision of 99% for the future. The 15-minutes-window as tolerated deviation to scheduled arrival times is often quoted, however a set of further definitions based on various views exists. The stipulation of a harmonized on-time performance term is additionally impacted by overlapping fields of responsibility in the air traffic system. Thus, a delay of planned processes within a functional range easily induces delays by overload in one or more sectors. These net-wide effects thereby crucially characterize the air traffic system as a dynamic one in terms of the (unwanted) plan variability.

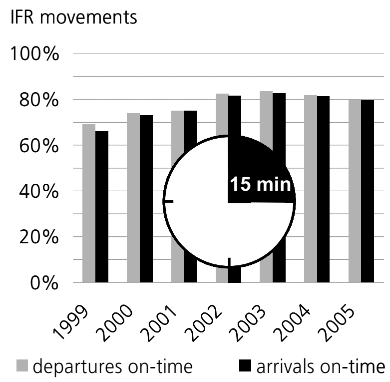

FIG 2. Punctuality situation in Europe
Data source: PRC [1]; Early dep/arr = on time

2. PRIMARY RESOURCE ALLOCATION

Commercial aviation requires a resource allocation phase for the net nodes – the airports – in order to determine the resources available as well as the actual demand. The potential capacity is split in minimum universal units and constitutes the infrastructure available as quantity of slots. A typical capacity unit is for example the maximum number of take-off's or landings per time-interval, often referred to as 10, 30 and 60 minutes. The daily to seasonal total traffic rate is to be defined in case of capping negative effects by emission of noise as well as exhaust gases.

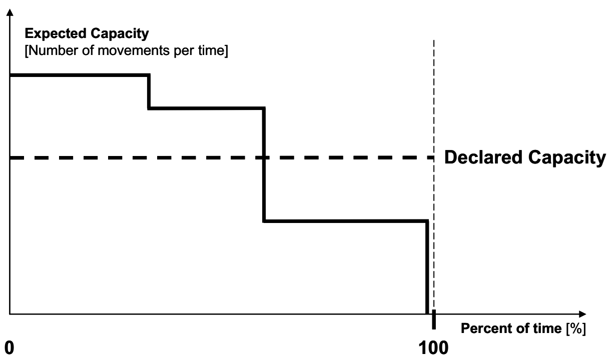

FIG 3. Capacity coverage chart

The major problem when defining the capacity potential is the extent approximation of the capacity range. However, the more accurate distribution characteristic of capacity data remains indistinct. Of course, empirical data are useful, but infrastructure is subject to a series of stochastically arising, capacity affecting environmental factors, like changing sights, wind speed and direction, just to mention a few. In FIG 3, a typical capacity trend is depicted. The abscissa is not scaled chronologically, but according to decreasing capacity.

The degree of effectiveness of the environmental factors determines the degree of difficulty to define the appropriate capacity rate, which allows acceptable performance rates, attainable in all conditions. On determination of the declared capacities usually it is accepted that the quality criterion (for example average delay per movement), already from a planning point of view, cannot be met at certain times. However, even when theoretically available (declared) capacity is higher than the expected demand, normal or abnormal operational deviations are likely to result in local bottlenecks. This is due to the fundamental relationship between utilisation of a resource and average delay per demander, as formulated in the queuing theory.

On the demand side, a process starts, which aims at the optimal harmonisation, with respect to the quantity of requests for slots and slots available, by shifting the desired times. In case of more requests than slots within a range of tolerance, many individual demands are to be excluded from the allocation until satisfactory allocation has been performed. At the end of this so-called primary allocation, each remaining slot-request will be met, albeit with deviations. Holding a slot implies the right to perform according flight movements at the time agreed upon. However, this is a very special right of use, since the holding airline is neither allowed to trade the right nor to insist on due compliance. Thus, the airline has to accept any operational lag without the right of compensation neither by the infrastructure service provider nor by the potential causer. On the other hand, the status quo does not allow any claim addressed to an airline causing lags for any third parties or extra cost for infrastructure service providers due to unscheduled resource usage. Furthermore, it should be noted, that the slot allocation authority is given to a coordinator by public appointment and not to the infrastructure service provider. Therefore, from a legal point of view, an infrastructure operator will not compensate for any delayed resource supply in the present situation. The respective law has to be altered, in order to raise the value perception with respect to on-time resource supply. The approach of punctual, delayed or lagged resources has not been implemented so far, instead, punctuality is understood as the schedule-fidelity of the respective airline. Infrastructure schedule-fidelity affects the on-time performance of the means of transportation, but is not directly measured. Instead, infrastructure is only considered to an extent, which is relevant for its capacity or capacity utilisation. However, capacity utilisation only defines the balance between capacity available and capacity required within a time-interval. The ratio of demand and supply only gives qualitative information on the punctuality to be expected from a system fragment. Utilisation of less than 100 % does not at all correspond to absolute operational punctuality, regardless of the tolerance between punctuality and unpunctuality. Flight schedules usually show a granularity of five minutes, due to the poor predictability during scheduling. The detailed order of resource usage, for example accurate approach and take-off sequences or even trajectories are not scheduled in advance, respectively the operational feasibility is not investigated.

The actual arrival distribution at the resource determines the extent of (un-) punctuality for each element. The value of the actually available capacity (not the targeted figure „declared capacity") is of same importance. The applicable separation minima (radar separation), the actually operated airplanes (wake vortex separation, speeds) as well as several relevant external factors determine this capacity.

3. SECONDARY RESOURCE ALLOCATION

The question about the allocation of capacity is posed every day again, but with higher granularity than during the scheduling phase. All the former resource allocation agreements, including the airport slot allocation, become relatively obsolete or are of minor importance, if a controller decides ad hoc about next actions. The prearrangements do only guarantee that schedules are theoretically realizable, at least regarding to mean values, with the expected quality of service.

Airplanes are generally serviced according to their arrival time at the requested infrastructure. The simple principle of First-Come-First-Served (FCFS) applies. This system is considered to act as a guarantor for fairness due to an equal treatment. Because of this equality, discrimination appears impossible. Furthermore, the strategy FCFS is characterized by simplicity because the next decisions in the context of aircraft sequencing are pre-determined, no decision making is required. Exceptions exist in operation in the case of an emergency or under any other abnormal circumstances. These special conditions are excluded from the proposed Level-of-Service-Concept.

The allocated rights to make use of scarce resources are valid even if the demanding aircraft is too early or late with reference to its slot or scheduled time (if no slot allocation process took place for this resource). The system FCFS ignores the risk of earning multiplied secondary delays. These delays occur, if the unpunctual demander will be inserted into any flow operating near the maximum throughput. The delayed demander disarranges the "following" aircraft, notwithstanding their current grade of punctuality. This re-sequencing usually results in additionally delays for the aircraft concerned – even though they were requesting the resource as allocated.

Aircraft causing alterations to the schedules of third parties got under the regime of FCFS no incentive to change this circumstance. As per the Central Office for Delay Analysis (CODA), airlines are responsible for about 50% of all primary delays in average (see FIG 4). This partition of delays is directly influenceable by the airlines; therefore, incentives to avoid violations of agreed schedules should become effective. These incentives would bring optimized usage of scarce resources due to reduced delays.

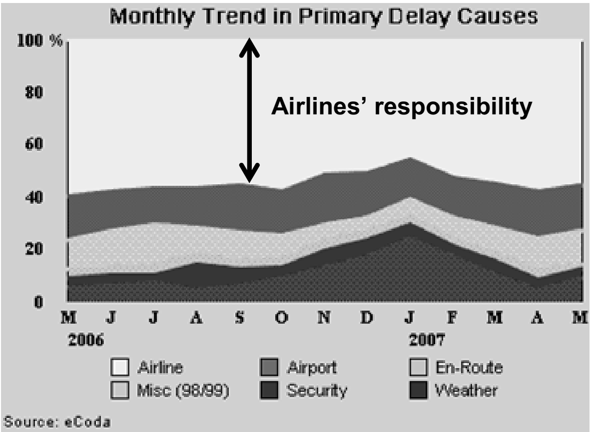

FIG 4. Primary delay causes
Source: eCODA

Overload due to increased demand happens, when primary delays trigger a secondary delay wave. Then flight movements (scheduled to the earlier intervals) would clash with those being on schedule in the sector observed. Since unused resources were temporally fixed before, they were not available any longer. The total demand for limiting resources could exceed the declared capacity and/or the usable capacity thereby. As mentioned before, the arising imbalance between capacity and demand is transferred by FCFS to all infrastructure users equally. With regard to the analysis of primary delay causes published by eCODA, according to which the airlines contribute significantly to the generation of delays, it is necessary to critically analyze delay propagation (even after FCFS). It seems impossible due to the complexity of infrastructure networks to strictly apply a costs-by-cause remuneration. However, the transmission of delays on stakeholders not involved could be absorbed. This would be an incentive to avoiding delays.

In the section before, main topic was how to solve the problem of partial overload situations due to additional (unpunctual) traffic. But there is another effect that is not tackled very well by FCFS; meant is here the situation of reduced capacities. In the case of reduced capacities at the day of operation, all the scheduled aircraft will come nevertheless. With FCFS, the remaining capacity will be re-allocated in chronological First-Come-First-Served order. Depending of the position in the waiting queue, an additional delay is imposed to each user. This delay increases while capacity remains reduced with later positions in the new sequence. The allocation of reduced capacities bases only on the sequence of requests; individual attributes do not play a role, although these attributes result from different economic utility functions.

More opportunities to optimise the sequencing strategy has an airline that dominates the resource, for example the home carrier at its hub. This airline can swap the slots of its fleet internally. The more dominant an airline is, the better economic chances are present. Every flight is endowed with a unique utility, based on the (individually defined) whole purpose of the mission. This utility function may have a special validity period and the benefit would change out of this range. Because of this, it should be clear, why a position shift in queueing systems can become very attractive. Today's strategy of the Air Traffic Control (ATC) does not care about the opportunities of the aviation stakeholders, which could result in increased connectivity, punctuality or any else, at least in greater dimensions than done in arrival and departure management systems. The occurrence of situations of reduced capacities (which is stochastically the case) can be seen as an external effect for the airlines concerned. The risk of being concerned by delay due to a lack of capacity is addressed to the airlines in general. This is one reason why to use buffer times in schedules, but buffer times reduce utilisation of personnel and aircraft. Additionally, the implementation of buffer times tends to result in early arrivals if capacity conditions were not decreased and thus the time buffer was not (fully) used up. To claim for an resource to early will lead to various problems with availability and rescheduling, if the resource is scarce and still used by other scheduled airlines.

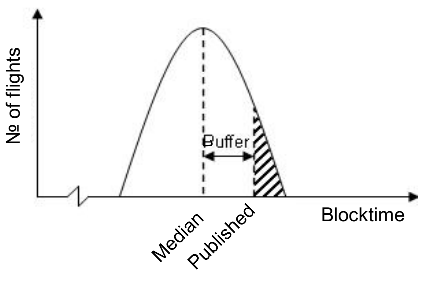

FIG 5. Buffer times in schedules

Source: Depicted according to [1]

4. LEVEL OF SERVICE CONCEPT

The search for an alternative procedure for resource allocation within the strategic and tactical range does not serve any end in itself, but the improvement of actual and future traffic quality. Regardless of the method to be found, it has to be measured on the criteria of economic sciences. It has to be proven, that the new method is the appropriate tool for socio-economic (or at least air traffic-economic) enhancement. Since the new procedure entails the change of the common distribution of resources, it can be seen as a change of the economic situation in the (air traffic) market and thus serves the tools of the welfare theory.

4.1. Pareto

An approach of the welfare theory (according to [2]) is analysing an economic change of a situation (e.g. alternative distribution of a quantity of goods) by means of the Pareto criterion. According to the Pareto-Criterion, an increase of the welfare is achieved, if the individual welfare of at least one subject increases, at the same time, however, no other subject experiences an individual welfare decrease. Since the change of economic situations mostly has winners and losers, the strict interpretation of the Pareto-Criterion is an argument against the change of status quo. From a mathematical point of view, this is a multi-criteria-related optimization problem, as known from the Operations Research (OR). Correspondingly, Pareto optimum exists, if the solution found can only be made accessible to one of the partial aims by rendering at least one aim that is more partial less accessible.

In terms of air traffic aspects, as contemplated here, the change of economic conditions means re-distribution of useable resources among demanders (subjects). Actually, all implementations in terms of quantity and quality of traffic management are mainly oriented towards the Pareto criterion, i.e., capacity-enhancing innovations are normally based on technical improvements (e.g. reduced separation minima) or capacity extensions (e.g. new runways), which serves all demanders by the equal treatment principle (Equality). Until now, the implementation of procedures, involving allocation priorities, has been hardly realised, due to the Pareto-conform necessity of individuals' disadvantaging.

4.2. Kaldor-Hicks

As described before, the strict interpretation of the Pareto-Criterion often prevents the implementation of alternative allocation mechanisms as to resource distributions. According to the Kaldor-Hicks-Criterion [2], nevertheless, this barrier might be overcome by the feasibility of fully compensating the disadvantaged client of the alternative distribution system while increasing social welfare anyhow. The heterogeneous distribution of the individual benefit acquired by a good is a pre-condition.

In terms of air traffic this means that individuals (airlines), who can acquire a higher benefit from the supply of scarce resources, even could acquire a higher benefit, if they would compensate the demanders, who were willing to refrain (those, who have less usage potential for the respective resource) according to their potential. The application of this way of compensation in fact results in the observance of the Pareto-Criterion. However, the circle of traded goods (slots) will be incremented by the quantity of the compensations. Potential compensations could be value-equivalent, e.g. money, or a value-balancing exchange of slots.

The above definition in this form helps to create a market-based approach for primary allocation schemes. In this paper, the focus lies on the question, how to distinguish infrastructure services, and how to compensate users for any loss of service quality?

4.3. Prepaid Priority

Instead of simply dividing the estimated available capacity into equivalent time niches (Slots), a certain priority is assigned to the slot now. Part of the available capacity (expected value), is assigned in form of highly prioritized slots (here called "A-Slots"), the following parts with lower priority/each. Priority means, that one user is preferred to another with lower priority for the use of the same scarce resource. In case of a conflict, the sequence of resource usage is pre-determined by the strategic assignment of a certain priority. The higher the priority of a flight, the less it is affected by resource scarceness. The less prioritized demanders compensate this effect by increasing waiting times on their side.

The average waiting time level of the infrastructure concerned - apart from losses due to technical difficulties - remains the same. The variances of process times in favour of high-prioritized slots are re-distributed. For the purpose of efficiency distinction, it is mandatory to assign all higher prioritized slots first, before allocating the subsequently lower category.

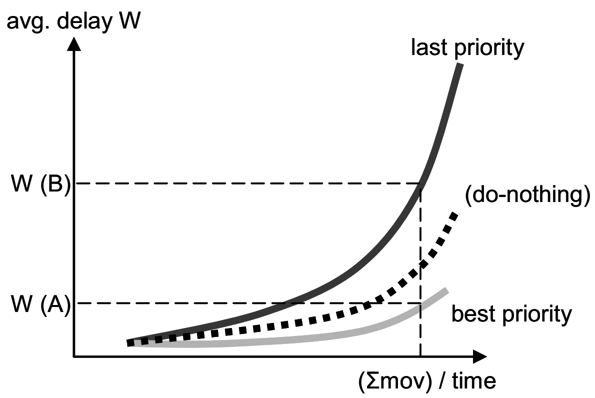

FIG 6. Splitting of the Demand-Delay-Curve run

If the demand does not exceed the capacity on the day of performance, the possession of a certain slot type is not relevant. In the case of prioritization coming into effect, remaining resources are no longer equally distributed to the demanders, but are made available to those, holding slots with the highest prioritization categories. Then, those holding slots with the lowest category almost completely bear the loss of capacity. The probability to be affected by resource scarceness thus depends on the individuals' readiness to pay.

4.4. Performance-based Priority

As outlined before, it seems to be essential to absorb the transmission of secondary delay waves to third parties. Such absorption could be realized, as on-time performance normally has priority to unpunctuality. An assigned slot, which has not been used at the arranged time, becomes obsolete. Only if a new slot is or becomes available, the unscheduled requesting aircraft can be served again. The absence of an automatic sequencing system for late (& early) traffic, enabling an aircraft, operating out of schedule, to utilize a scarce resource anyway by suppressing punctual operating scheduled traffic, generally increases the motivation to operate punctually. This thesis is based on the airlines' joint responsibility for primary delays and the resulting, surely not yet entirely used potential of action, to comply with operation times. A punctually operating airline, which uses time buffers for on-time performance optimization, can reduce buffer times by acquiring high prioritized slots, since a substantially reduced variation of the block times may the result.

4.5. Servicelevel-based Resource Allocation

By combination of the models described before, scheduling and operations in particular shall become stable and transparent. Thus, prioritizing allocation by means of distinct slot-types is to be agreed upon in advance, in the event of increasing scarceness of resources, according to the airlines' requirements and net stability. Simultaneously, a waste of resources due to unpunctuality is to be reduced.

In such a Level-of-Service-system, all aircrafts would gain resource usage rights by strictly referring to their priority. At the same time, the characteristic "priority" depends on the characteristic "on-time performance". Thus, an aircraft loses its priority by unpunctuality.

Assuming that in our model exist two categories of prioritisation A and B - whereas A implies highest prioritisation - which can be pre-selected by the airlines, different priorities will result regarding to the conditions in the case of infrastructure scarceness. To give an incentive to perform flights as punctual as possible, the pre-selected priority class (A or B) is only guaranteed when requesting the scarce infrastructure within the allocated punctuality window. Otherwise the selected prioritisation class will be downgraded, A to "α" and B to "ß". The complete prioritisation order looks like this: A-B- α- ß. This method of resource allocation when controlling the flow is as fair as possible due to the independency of unpunctual traffic.

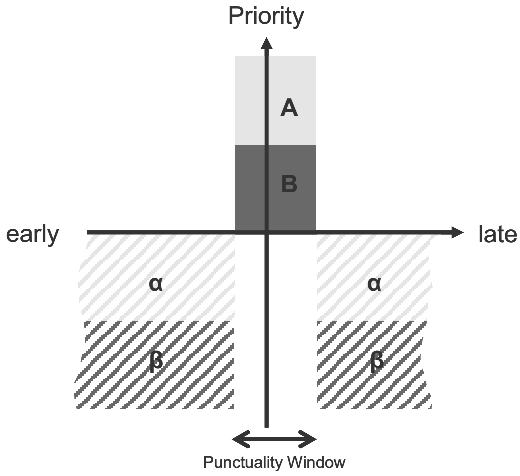

FIG 7. Individual priority subjected to pre-paid priority and performance

In addition to A-B-α-β is A-α-B-β applicable. However, this variant might encounter an extremely one-sided interest, due to the fundamental preference of financially strong customers. Additionally, the aim to punctual operation is evidently smaller and the capacity-increasing objective of capacity waste avoidance is failed.

Fundamental basis of the performance-based priority is the existence of a consistent on-time performance definition. The 15-minute-regulation is only considering the difference between published and reached on block time and does not make any statement about intermediate on-time departure punctuality or about turnaround-punctuality (time between on block and off block). However, for a successful implementation it would be mandatory to introduce harmonized punctuality standards in every operational sector (according to the responsibility of the service provider: air space sectors, runway, apron, gate), as well as for the parties serviced. In order to ensure the optimal compliance with these standards, penalty clauses should be implemented into the respective service level agreements.

5. DELPHI-STUDY ABOUT DISTINGUISHED INFRASTRUCTURE SERVICE LEVELS

5.1. The Methodology of the Delphi-Survey

The Delphi-method is one of the most popular scientific forecast tools. The name is based on the 2800 year ago founded oracle of Delphi. Believing the legend, on requests about the future, the woman priest Pythia – as a mouthpiece of Apollo – announced a kind of wisdom and this wisdom showed the answer. The methodology of the here used Delphi was created by RAND Corporation.

The basic principle of this technique is to repeatedly consult an expert panel about nebulous developments, trends and forecasts. Therefore, feedback about the panel's estimations of the last turn is provided to each participant. Instead of face-to-face-interviews or panel discussions, which can be very cost-intensive and time-consuming, questionnaires will be provided. The Delphi technique is a brilliant tool to produce high quality results, especially in comparison to other techniques. Special advantage results due to the anonymous questioning situation. With that, unwanted group dynamic processes can be eliminated, which would falsify the results. Additionally, the variability of opinions will be decreased during the follow-up phase, because experts with isolated opinions based more on guesswork than knowledge get the chance to rework the answers. On the other hand, experts are free to validate the given answers.

5.2. Intermediate Results

When editing this paper, the Delphi-Study with the title „Level of Service for Aviation Infrastructure Markets?“ was still in progress, but some results out of the first wave can be taken to give an introduction. The Delphi survey is prepared both as paper-pencil-questionnaire and online questioning. All versions are available in English and German language. The panel consists of 52 aviation researchers, practitioners, regulators and politicians. One half of the panellists work in Germany, altogether experts from 14 countries take part.

For the future, most of the experts expect service quality to be much more important, after a decade of special price-sensitiveness. For commercial air traffic, punctuality is announced to become more important (FIG 8).

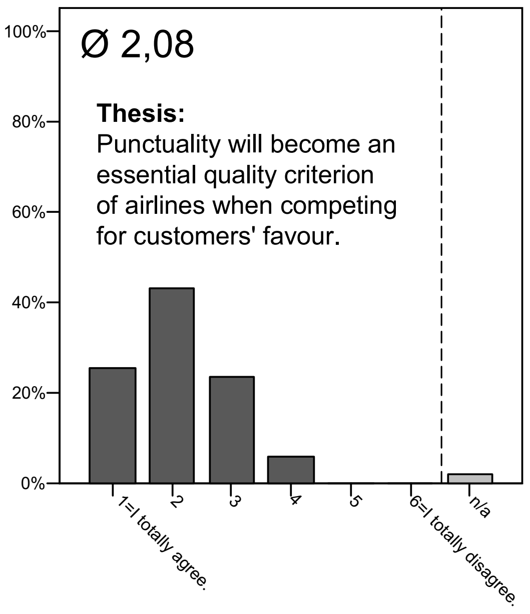

FIG 8. Response about future quality issues

All experts believe that airlines have an influence on punctuality, and more than 2/3 of the panel see even a (rather) great potential to improve on-time performance (FIG 9). This statement correlates with the statistics of Eurocontrol/eCODA, which call the airlines responsible for about 50% of all primary delays.

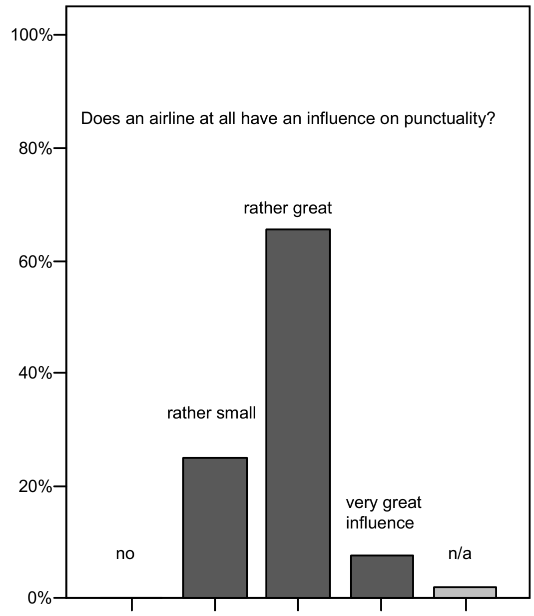

FIG 9. Response about airline’s influence on punctuality

The majority of the panellists expect a change in the primary airport slot allocation scheme to become effective. The administrative distribution will be replaced by an economically oriented system.

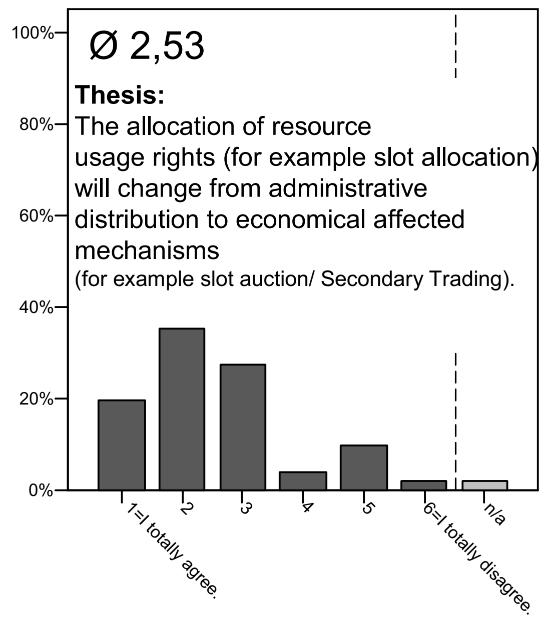

FIG 10. Response about future resource allocation

Most interviewed experts prefer Equity to become the basic principle in the context of resource allocation. Individual attributes would play a role in this case, however, only a smaller part patronises a strictly equal treatment.

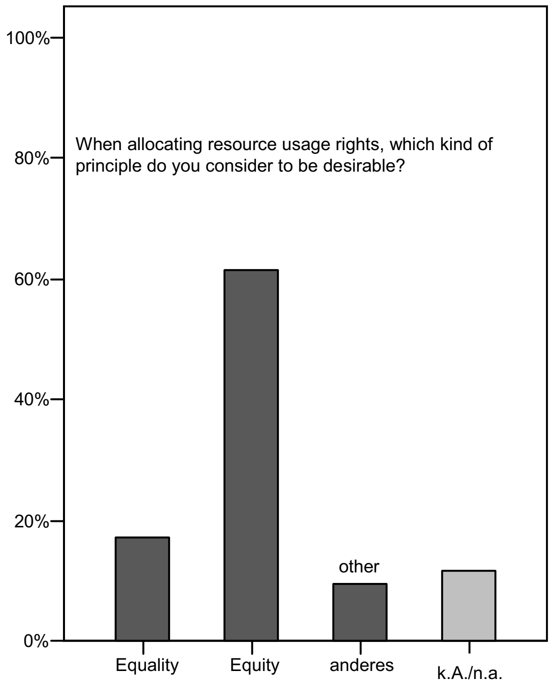

FIG 11. Response about the favoured allocation principle

6. CONCLUSION

The resource allocation process described combines for the first time strategic and tactical planning processes. It smoothes the net-wide impacts of bottle-neck situations according to the originator and encourages economic-oriented usage of resources. In this model the airlines are left to decide either for expensive prioritization with the resulting economical potential by reduction of block-time variances or the less expensive “B-rights”, which also can generate costs by necessary buffer dimensioning. Until now, this chance of trade-off does not exist in the air transport system, but could be implemented for the purpose of economic-oriented product categorizing in the future. The primary allocation method is of lower importance. It is appropriate for the European status quo, but also for the discussed alternative way of primary allocation distribution, involving amongst others slot trading rights, auctions or lotteries.

When analysing the customer’s reaction in the event of application, the legal question of competition-related criteria of the allocation process has to be considered. It may not seem to be unrealistic, that some airlines will show a particularly high readiness to pay for highly prioritized slots at their hubs and thus complicating the market entrance for small companies by the high price level in this segment. Simultaneously, this very chance of choosing service levels is a chance for newcomers for low priced slot allocation at all. Furthermore, the described slot allocation strategy is compatible with other ideas of air traffic system improvement, like the reservation of certain trajectories (gateways) for highly frequented origin-destination-flows, which also can be used for the implementation of certain quality aims.

Future research has to analyze, which type of distinction is appropriate. It has to be clarified, how many prioritization categories are reasonable and which kind of proportion. Furthermore, the frame conditions’ parameters, appropriate to meet a future service level agreement, have to be stipulated. These are mainly the prohibited deviations (without penalty) from the aimed value (definition “punctuality”) and the penalty regulations in case of failure, but additionally financial estimations of the service level fair market values. Economic aspects, aiming to enhancement of social benefit, should replace individual optimization requests (e.g. for maximum aircraft flows). The aim should be a maximisation of added value along the entire travel chain (at least the air transportation segment). However, correlation of a maximum demander-flow is not mandatory, since, due to this model, it is dispersed more heterogeneous in terms of value than it is right now. Furthermore, the model should be investigated in terms of economic advantages compared to other strategies, e.g. status quo when allocating slots and regular operational traffic flow control according to “First Planned First Served” (FPFS). Finally, the question of net-wide service level implementation is still open. Which kind of service levels are valid for airports, are they distinct or is there a need of homogeneity?

References

[1] Eurocontrol, Performance Review Commission: Performance Review Report 2005. Brussels, Belgium 2006
[2] Luckenbach, H.: Theoretische Grundlage der Wirtschaftspolitik. Verlag Vahlen, München, 2000
[3] Häder, M.: Delphi-Befragungen. Ein Arbeitsbuch. Westdeutscher Verlag, Wiesbaden, 2002